Using Two Blinking Eyes and One Talking Mouth for Fast Face Location with Complex Backgrounds in Internet or Video

CHIUNHSIUN LIN
National Taipei University
69 Sec. 2 Chian_Kwo N. Road, Taipei, Taiwan, 10433
TAIWAN

Abstract: - Automatic human face location system that uses two blinking eyes and one talking mouth to locate the only one human face embedded in Internet or video images is presented. The designed system is composed of two principal parts: The first part is to detect the potential face regions that are obtained from the criteria of "the combination of the two blinking eyes and one talking mouth". The second part of the proposed system is to perform the face verification task by using a support vector machine (SVM) classifier. The experimental results reveal that the proposed method is outstanding in terms of efficiency and accuracy.

Key-Words: - Face location; Internet; Video; Motion; Support vector machine, Internet security

1 Introduction

Automatic locating of human faces is the first step for face recognition or face identification, and it can be used as ″login process″ in Internet security, E-commerce safety, and ATM (Automatic Teller Machine). However, locating and tracking a human face robustly in real time is still a tough problem. The key distinction between face localization and face detection centers on the number of faces in an image. Face localization is to locate the only one human face embedded in an image. Face detection is to detect one or more human faces embedded in an image. Some successful systems have been proposed in the literature, such as [1-16]. Rowley [15] presented a neural network-based face detection system by using a retinal connected neural network to check the small windows of an image, and judge whether each window contains a face or not. They adopted a small window (20 * 20) to slide over all portions of an image at various scales and used oval mask for ignoring background pixels. Their system arbitrates between multiple networks to improve the performance over a single network. However, the inefficient search is a time-consuming procedure. Lin [16] presented a triangle-based approach system that can locate multiple faces embedded in complicated backgrounds. Moreover, it is able to handle different size, different lighting condition, varying pose and expression. However, when the condition is complicated, their system becomes slow and time-consuming. Therefore, we utilize two blinking eyes and one talking mouth to accelerate the executing time of locating human face in complicated backgrounds case.

In the investigation, we present a system that can handle different size, different lighting condition, varying pose and expression at the same time. The overview of our system is shown in Fig. 1. The first part of the designed system contains three phases. First, read in a set of frames/images from Internet or video, and then obtain the difference between two frames (e.g. frame 1 and frame 2). Second, convert the difference frames/image to a binary image. Label all 4-connected components in the image and count the number of the blocks. Third, detect any three centers of three different blocks constitute two blinking eyes and one talking mouth. Subsequently, clip the blocks that satisfy the two blinking eyes and one talking mouth criteria as the potential face region. In the second part, each face candidate obtained from the previous process is resized to a standard size. Next, each of these normalized potential face regions is fed to the SVM classifier to verify whether the potential face region really contains a face or not. The rest of the paper is organized as follows. In section 2, Searching for potential faces are illustrated. In section 3, each of the normalized potential face regions is fed to the SVM classifier to verify whether the potential face region really contains a face or not. Experimental results are demonstrated in section 4. Finally, conclusions are given in section 5

2 Searching for Potential Face Regions

In real application (e.g. MSN Web Messenger), we can ask the testers/users to blink their eyes and talk on purpose, so we assume the testers/users will blink their eyes and talk frequently. The task of face location can be divided into the following steps. First, read in a set of frames/images from Internet or video, and then obtain the difference between two frames (e.g. frame 1 and frame 2). Second, using threshold = 80 to convert the frames/image of difference to a satisfied binary image and remove some noise (the threshold could be 80~100, and make almost no difference). Third, label all 4-connected components in the image to form blocks and count the number of the blocks. If ″the number of the blocks″ is equal to 3 or larger than 3, then we will find out all the centers of all blocks. Otherwise, it will go to the loop of checking "the difference between two frames (e.g. frame 1 and frame 3)." For keeping away from infinite loop, we use a variable as "count", if the variable "count" is equal to 20, then the program will go to the next phase (find out any three centers of three different blocks that satisfy "the combination of two eyes and one mouth (isosceles triangle)". Algorithm for obtaining the satisfied binary image of the frame of difference is described as follows:

Step (0) Initiate the count = 0 (first loop).
Step (1) Read in a set of frames/images from Internet or video, and then obtain the difference between two frames (e.g. frame 1 and frame 2).
Step (2) Using threshold = 80 to convert the frames/image of difference to a satisfied binary image and remove some noise (the threshold could be 80~100, and make almost no difference). Next, label all 4-connected components in the image and count the number of the blocks.
Step (3) If the number (N) of the blocks is equal to 3 or larger than 3 or count = 20, stop the loop, and then go to the second phase (find out all the centers of all blocks).
Step (4) If N < 3, count = count + 1, then goes to the first step to detect any two blinking eyes and one talking mouth by the difference between the other two frames (e.g. frame 1 and frame 3; then frame 1 and frame 4 …; frame 1 and frame 21).

Fourth, find out all centers of all blocks, and then find out any three centers of three different blocks that satisfy an isosceles triangle that contains two blinking eyes and one talking mouth in the human face will form. If the triangle i j k is an isosceles triangle as shown in Fig. 2(a), then it should possess the characteristic of “the distance of line i j = the distance of line j k”. From observation, we discover that the Euclidean distance between two eyes (line i k) is about 90% to 110% of the Euclidean distance between the center of the right/left eye and the mouth. Due to the imaging effect and imperfect binarization result, a 25% deviation is given to absore the tolerance. Here, “abs” means the absolute value, “D (i, j)” denotes the Euclidean distance between the centers of block i (right eye) and block j (mouth), “D (j, k)” denotes the Euclidean distance between the center of block k (left eye) and block j (mouth), “D (i, k)” represents the Euclidean distance between the centers of block i (right eye) and block k (left eye).The first matching rule can thereby be stated as (abs(D(i, j)-D(j, k)) < 0.25*max(D(i, j), D(j, k))), and the second matching rule is (abs(D(i, j)-D(i, k)) < 0.25*max(D(i, j), D(j, k))). Since the labeling process is operated from left to right then from top to bottom, we can get the third matching rule as “i < j < k”. For example, as shown in Fig. 2(b), if three points (i, j, and k) satisfy the matching rules, then we think that they form an isosceles triangle. After we have found the two blinking eyes and one talking mouth, we use the Euclidean distance between the centers of block i (right eye) and block j (left eye) to get the coordinates of the four corner points that form the potential facial region as shown in Fig. 3.

Fifth, clip the blocks that satisfy the triangle criteria as the potential face regions. Since we think that the real facial region should cover the eyebrows, two eyes, mouth and some area below the mouth, the coordinates can be calculated as follows:

$$X1 = X4 = Xi - 1/4*D\ (i, j); \quad (1)$$
$$X2 = X3 = Xj+1/4*D\ (i, j); \quad (2)$$
$$Y1 = Y2 = Yi+1/4*D\ (i, j); \quad (3)$$
$$Y3 = Y4 = Yj -1/4*D\ (i, j); \quad (4)$$

Finally, we normalize all of the potential face regions to a standard size (30 * 30 pixels). Here, we resize the potential face region using "bicubic" interpolation technique. Herein, the potential facial region is resized by the "bicubic" interpolation technique as described in the textbook written by Gonzalez et al. [17].

3 Face Verification

In the previous section, a set of normalized potential face regions (30 * 30 pixels) in an image was selected. This section presents an efficient support vector machine classifier to determine whether a potential face region contains a face or not.

The SVM is a new and promising classification and regression technique proposed by Vapnik and his group at AT&T Bell Laboratories. The main idea of SVM comes from (1) a nonlinear mapping of the input space to a high dimensional feature space, and (2) given two linearly separable classes, designs the classifier that leaves the maximum margin from both classes in the feature space. SVM, which displays good generalization performance, has been applied extensively for pattern classification, handwriting recognition, data mining, regression, and density estimation.

Let x i, i = 1, 2 . . . , N, be the feature vectors of the training set, X. These belong to either of two classes, ω1, ω2, which are assumed to be linearly separable. The goal is to design the decision hyperplane:

$$g(\underline{x}) = \underline{w}^T \underline{x} + w_0 = 0 = w_1 x_1 + w_2 x_2 + ... + w_l x_l + w_0 \qquad (5)$$

Assume $\underline{x}_1, \underline{x}_2$ on the decision hyperplane :

$$0 = \underline{w}^T \underline{x}_1 + w_0 = \underline{w}^T \underline{x}_2 + w_0$$

$$\Rightarrow \underline{w}^T (\underline{x}_1 - \underline{x}_2) = 0 \;\; \forall \underline{x}_1, \underline{x}_2 \qquad (6)$$

Since the difference vector x1 –x2 obviously lies on the decision hyperplane (for any $x_1 - x_2$), it is apparent from equation (6) that the vector ω is orthogonal to the decision hyperplane. The decision hyperplane is also called Optimal Separating Hyperplane (OSH) that minimizes the risk of misclassifying not only the samples in the training set but also the other samples of the test set. For a two class classification problem, the goal is to separate the two classes by the only decision hyperplane that is established by available samples. Consider the samples in Fig. 4(a), where there are many possible linear classifiers (e.g. H1, H2, H3 and H4) that can separate the data, but there is only one decision hyperplane (shown in Fig. 4(b)) that maximizes the margin (the distance between the decision hyperplane and the nearest data point of the two classes). SVM has similar origins with neural networks, but recently it has been widely used in classification. SVM used some statistical learning theory to solve these problems in reasonable time. The software of SVM that we used can be obtained from: "http://www.csie.ntu.edu.tw/%7Ecjlin/", and SVM is described in detail in the textbook written by Theodoridis et al. [18].

The support vector machine classifier that we used was trained by using a large number of face and nonface examples. In order to locate faces at a range of rotation, non-uniform illumination and various expressions, we also incorporate the faces of these categories into our face training database. Each normalized potential facial region is fed into the support vector machine classifier to perform the verification. Once a face region has been confirmed, the final step is to remove regions that overlap with the selected face region, and then output the result.

4 Experimental Results

In this section, a set of experimental results is demonstrated to verify the effectiveness and efficiency of the proposed system. In real application (e.g. MSN Web Messenger), we can ask the testers/users to blink their eyes and talk on purpose, so we assume the testers/users will blink their eyes and talk frequently. Therefore, we use a database containing 30 different persons is taken from Internet or video in the assumption (the testers/users will blink their eyes frequently). One example of our experimental processing demonstrates step by step as shown in Fig. 5. Fig. 5(a) displays frame 1; Fig. 5(b) illustrates frame 2: Fig. 5(c) shows the frame/image of difference between frame 1 and frame 2: Fig. 5(d) depicts the satisfied binary image of Fig. 5(c); Fig 5. (e) demonstrates two blinking eyes and one talking mouth constitute an isosceles triangle; Fig 5. (f) shows the normalized original potential facial region (30 * 30 pixels) that is fed into the support vector machine classifier to perform the verification; Fig 5. (g) depicts the final result. The resolution of each frame/image is 240*180 pixels. Experimental results demonstrate that a 100% success rate is achieved.

One example of Lin [16] experimental processing demonstrates step by step as shown in Fig. 6. Fig. 6(a) displays the input image; Fig. 6(b): depicts the satisfied binary image of Fig. 6(a); Fig. 6(c) shows two eyes and one mouth constitute an isosceles triangle; Fig. 6(d) demonstrates the binary potential facial region that is based on two eyes and one mouth; Fig. 6(e) displays the original potential facial region that is based on two eyes and one mouth; Fig. 6(f) shows the binary potential facial region (60*60 pixels) that is fed into the weighting mask function to perform the verification;; Fig. 6 (g) depicts the final result. From Fig. 5 and Fig. 6, we depict two systems with/without detection of two blinking eyes and one talking mouth constitute an isosceles triangle in complex backgrounds. Fig. 5(d) depicts the satisfied binary image with detection of two blinking eyes and one talking mouth constitute an isosceles triangle, and the number of blocks is only 12; Fig. 6(b) by using the scheme without detection of two blinking eyes and one talking mouth in Lin [16], and the number of blocks is 51; Due to the decrease of the number of blocks (from 51 to 12), we can expect that the speed will be improved drastically in the complicated background case. Fig. 6(a) depicts a images with 240*180 pixels, and it needs 97.3594 seconds in Lin [16] (the number of blocks is 51) to locate the correct face position by using a P4 CPU 3.0 GHz PC. The same image needs only 1.2656 seconds to locate the correct face position in this proposed work. Moreover, the proposed scheme is significantly faster than Rowley [15], since they adopted a small window (20*20) to slide over all portions of an image at various scales is a time-consuming procedure.

5 Conclusion

We proposed an effective face location system that uses two blinking eyes and one talking mouth to accelerate the executing time of locating human face with complicated backgrounds case in Internet or video images. The experimental results reveal that the proposed method is outstanding in terms of efficiency and accuracy. Briefly, the proposed novel algorithm is significantly faster than previous investigations in Rowley [15] and Lin [16] in the case of complex backgrounds. Furthermore, if the testers/users fail to blink their eyes and talk, then our system still can locate the human face by using the triangle-based and SVM approach. In this case, our system will become slower certainly. In the future, we plan to extend our database and use this face location system as a preprocessing for solving face recognition problem.

References:

[1] V. Govindaraju, S. N. Srihari, D. B. Sher, "A computational model for face location", Proc. Computer Vision and Pattern Recognition, 1990, pp. 718-721.

[2] H. L. Choong, S. K. Jun, H. P. Kyu, "Automatic human face location in a complex background using motion and color information", Pattern Recognition Vol. 29, Issue: 11, 1996, pp. 1877-1889.

[3] Y. Dai, Y. Nakano, "Face-texture model based on SGLD and its application in face location in a color scene", Pattern Recognition, Vol. 29, no. 6, , 1996, pp. 1007-1017.

[4] P. Juell, R. Marsh, "A hierarchical neural network for human face location", Pattern Recognition, Vol. 29, no. 5, 1996, pp. 781-787.

[5] J. W. Kim, B. H. Kang, P. M. Kim, M. S. Cho, "Human face location in image sequences using genetic templates", IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, 1997, pp.2985 - 2988.

[6] L. S. Shen, K. Q. Wang, X. Xing, "Automatic human face location and tracing in a complex background", Chinese Journal of Electronics, Vol. 9, 2000, pp. 65-69.

[7] D. Maio, D. Maltoni, "Real-time face location on gray-scale static images", Pattern Recognition, Vol. 33, 2000, pp. 1525-1539.

[8] Y. J. Wang, B. Z. Yuan, "Robust face location and tracking using optical flow and genetic algorithms", Chinese Journal of Electronics, Vol. 10, 2001 pp. 450-454.

[9] Y. Wang, B. Yuan, "Fast method for face location and tracking by distributed behaviour-based agents", Vision, Image and Signal Processing, IEE Proceedings, Vol. 149, Issue 3, 2002, pp. 173 – 178.

[10] P. Sharma, R. Reilly, "Fast marching methods applied to face location in videophone applications using colour information",IEEE International Conference on Multimedia and Expo, Vol. 2, 2002, pp. 141 – 144.

[11] Tianxiang Yao, Hongdong Li, Guangyao Liu, Xiuqing Ye, Weikang Gu, Yiqing Jin, "A fast and robust face location and feature extraction system", International Conference on Image Processing, Vol. 1, 2002, pp. I-157 - I-160.

[12] J. S. Tang, S. Acton, "Locating human faces in a complex background including non-face skin colors", Journal of Electronic Imaging, Vol. 12, 2003, pp. 423-430.

[13] B. Raducanu, M. Grana, F. X. Albizuri, A. d'Anjou, "A probabilistic hit-and-miss transform for face localization", Pattern Analysis and Applications, Vol. 7, 2004, pp. 117-127.

[14] Wang Xianbao, Cao Wenming, Li Guojun, "A Method of Face Location in Complex Background", International Conference on Neural Networks and Brain, Vol. 3, 2005, pp. 1507 – 1510.

[15] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," IEEE Transactions on PAMI, Vol. 20, Issue 1, 1998, pp. 23 – 38

[16] Chiunhsiun Lin, Kuo-Chin Fan, "Triangle-based Approach to the Detection of Human Face," Pattern Recognition, Vol. 34, No. 6, 2001, pp. 1271-1284.

[17] Rafael C. Gonzalez and Richard E. Woods, "Digital Image Processing", Addison-Wesley Publishing Company, 1992.

[18] S. Theodoridis, K. Koutroumbas, "Pattern Recognition", 2nd edition, Elsevier Academic Press, 2003.

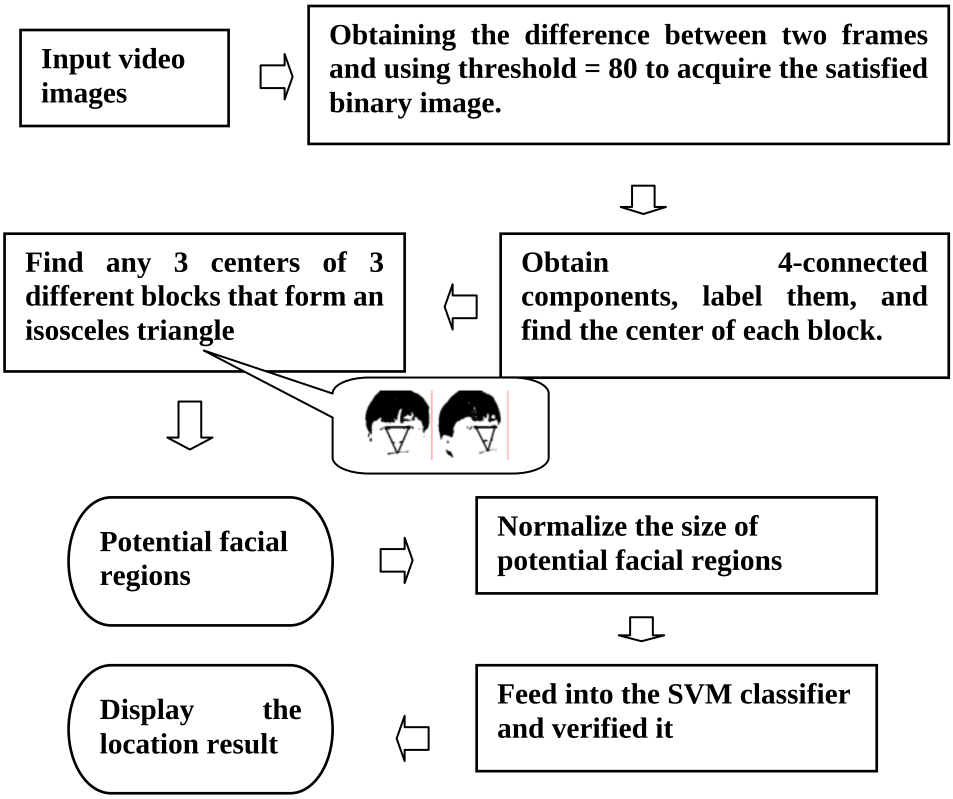
Input video images
Obtaining the difference between two frames and using threshold = 80 to acquire the satisfied binary image.
Obtain 4-connected components, label them, and find the center of each block.
Find any 3 centers of 3 different blocks that form an isosceles triangle
Potential facial regions
Normalize the size of potential facial regions
Feed into the SVM classifier and verified it
Display the location result

Fig. 1. Overview of our system

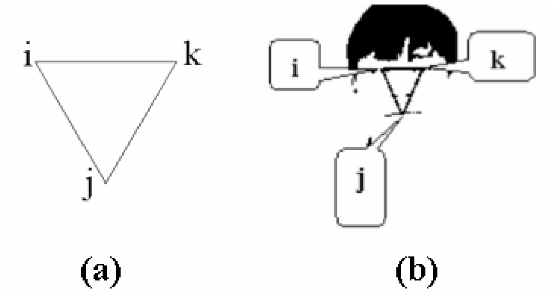

Fig. 2. The isosceles triangle i j k. (b) Three points (i, j, and k) satisfy the matching rules, which will form an isosceles triangle.

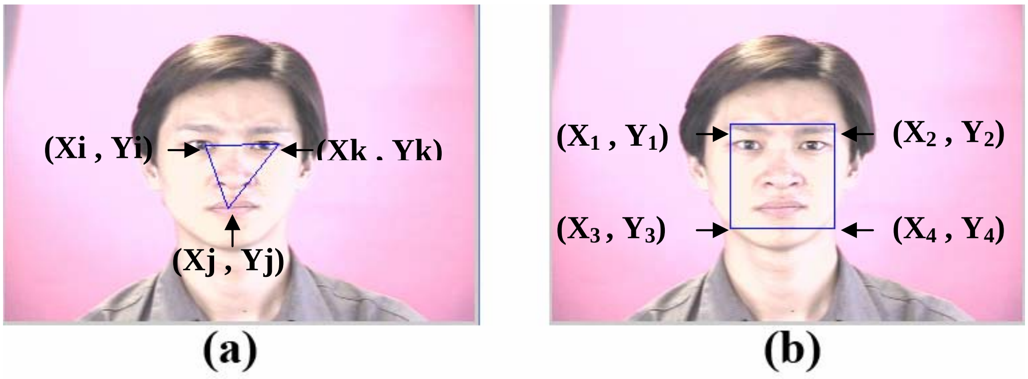

Fig. 3 IF (X*i*, Y*i*), (X*j*, Y*j*) and (X*k*, Y*k*) are the three center points of blocks i(right eye), j(mouth), and k(right eye), respectively. The four corner points of the face region are thus (X1, Y1), (X2, Y2), (X3, Y3), and (X4, Y4).

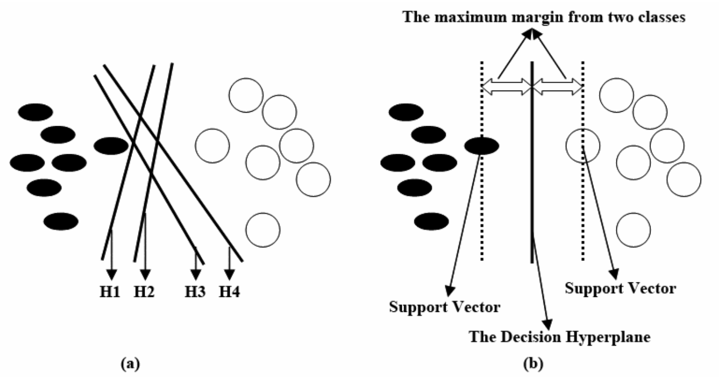

Fig. 4 The only one decision hyperplane maximizes the margin.

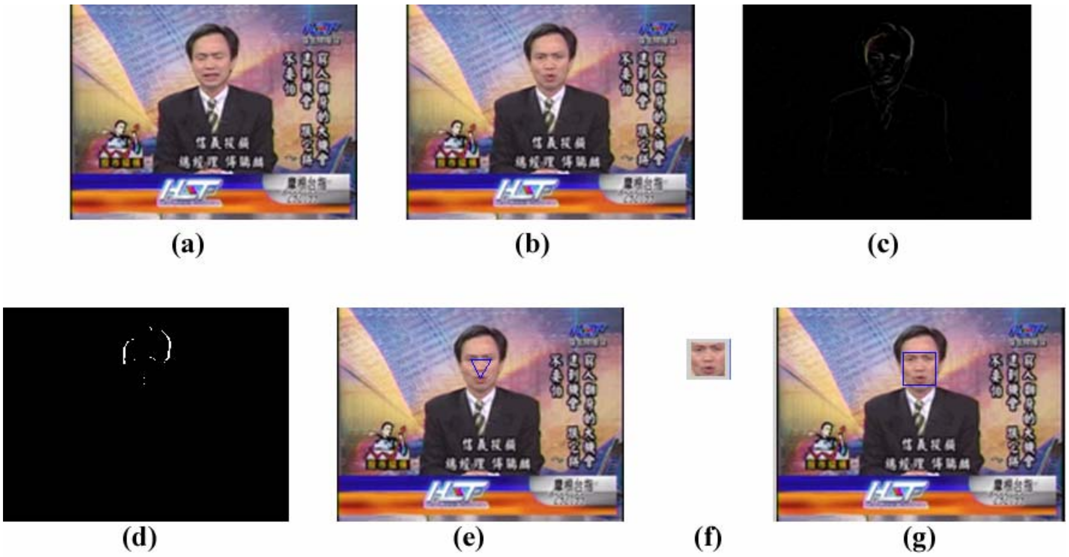

Fig. 5 Our experimental processing demonstrates step by step.

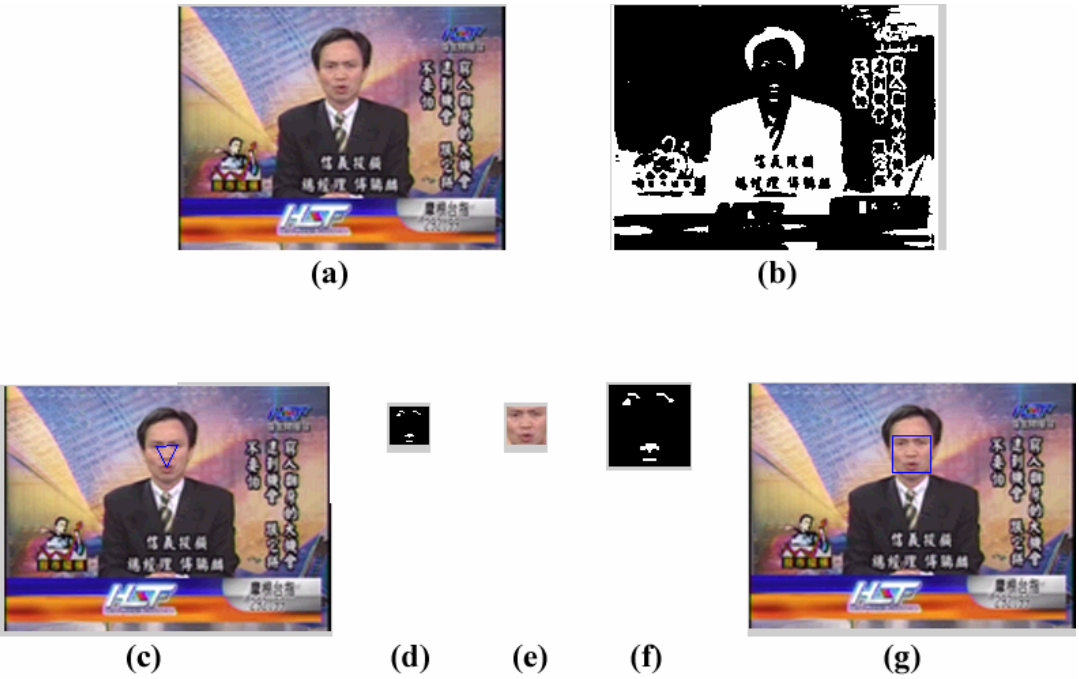

Fig. 6 Experimental processing of Lin [16] displays step by step.